



OPEN ACCESS

EDITED BY

Aslina Baharum,
Sunway University, Malaysia

REVIEWED BY

Arash Zaghi,
University of Connecticut, United States
Puti Xu,
Victoria University, Australia

*CORRESPONDENCE

Roberto Angel Melendez-Armenta
✉ ramelendeza@itsm.edu.mx

RECEIVED 28 October 2025

REVISED 08 January 2026

ACCEPTED 13 January 2026

PUBLISHED 05 February 2026

CITATION

Luna Chontal G, Melendez-Armenta RA,
Degante-Aguilar E and
Fernández-Domínguez FJ (2026) Task
automation and instructional planning
support with large language models: a
systematic review.
Front. Educ. 11:1733861.
doi: 10.3389/feduc.2026.1733861

COPYRIGHT

© 2026 Luna Chontal, Melendez-Armenta,
Degante-Aguilar and Fernández-Domínguez.
This is an open-access article distributed
under the terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other forums is
permitted, provided the original author(s) and
the copyright owner(s) are credited and that
the original publication in this journal is cited,
in accordance with accepted academic
practice. No use, distribution or reproduction
is permitted which does not comply with
these terms.

Task automation and instructional planning support with large language models: a systematic review

Giovanni Luna Chontal¹, Roberto Angel Melendez-Armenta^{1*},
Edgar Degante-Aguilar² and
Francisco Javier Fernández-Domínguez¹

¹Affective Computing and Educational Innovation Laboratory, Division of Graduate Studies and Research, Tecnológico Nacional de México-Instituto Tecnológico Superior de Misantla, Misantla, Veracruz, Mexico, ²Tecnológico Nacional de México-Instituto Tecnológico Superior de Teziutlán, Teziutlán, Puebla, Mexico

Introduction: The use of large language models (LLMs) in education has rapidly expanded, generating interest in their potential to support teachers through task automation and instructional planning. This review synthesizes evidence on reported changes in time devoted to educational content generation and support for planning-related task automation.

Methods: We conducted a systematic review of peer-reviewed literature published between 2023 and 2025, focusing primarily on secondary and higher-education contexts. Study selection followed PRISMA 2020 guidelines. Risk of bias was assessed using ROBINS-I for non-randomized studies and CASP checklists for qualitative/mixed-methods studies and for secondary evidence syntheses.

Results: Sixteen studies met inclusion criteria (13 primary empirical studies and 3 secondary syntheses). Across primary studies, LLM use was associated with reported time savings and perceived gains in clarity or usefulness of generated educational resources. However, outcomes and measures were heterogeneous and often self-reported, several risk-of-bias domains were rated as unclear, and evidence was concentrated in higher-education settings with small samples, limiting comparability and causal inference. Recurrent constraints included limited reproducibility, strong dependence on prompt design, and ethical or technical concerns.

Discussion: LLMs may support educators, but conclusions should be interpreted cautiously. Effective integration requires clear pedagogical frameworks, human oversight, and standardized evaluation in real-world settings.

Systematic review registration: Open Science Framework (OSF). Unique identifier: v63nj. Public URL: <https://osf.io/v63nj/>

KEYWORDS

artificial intelligence in education, automated feedback, instructional planning support, large language models, lesson design, systematic review, teacher workload

Introduction

The rapid advancement of digital technologies has clearly exposed the structural limitations of traditional educational systems in addressing 21st-century challenges such as personalized learning, equitable access to high-quality resources, and support for students with diverse needs (Çapuk and Kara, 2015; Makgato, 2014; Onyema, 2019). Despite widespread

acknowledgment of these issues, a significant gap remains in many developing countries between the availability of technological infrastructure and its effective integration into school contexts (Dotong et al., 2016; Jhurree, 2005). Recent studies suggest that this lack of integration is largely due to factors such as limited resources, institutional resistance to change, and insufficient teacher training in digital competencies (Fernández-Batanero et al., 2022; MacKinnon and MacKinnon, 2013). Furthermore, traditional teaching methodologies continue to prevail in many classrooms, restricting opportunities for pedagogical innovation and adaptation to students' needs (Ramorola, 2013). This situation has heightened the urgency to transform educational environments into more dynamic, adaptive, and inclusive models capable of meeting new standards of educational quality and sustainability (Mansur et al., 2023).

One of the most promising emerging alternatives for addressing these challenges is the use of Large Language Models (LLMs) artificial intelligence tools that are capable of generating personalized content, providing real-time pedagogical assistance, and fostering more interactive, context-aware learning experiences (Chen et al., 2024).

Large language models such as GPT-4 are built on deep neural networks trained on massive volumes of textual data, which allow them to grasp context, generate coherent text, and adapt their responses to the user's communicative style (Li et al., 2024; Sharma et al., 2025). These capabilities have been leveraged to create intelligent educational environments that support personalized learning, automated conversational tutoring, and immediate feedback (Huang et al., 2023; Ningsih and Lahby, 2025; Park et al., 2024). In medical education, for instance, LLMs have been successfully employed in virtual patient simulations to train clinical skills (Abd-Alrazaq et al., 2023), while in higher education they are increasingly integrated as pedagogical assistants and content generators (Alier et al., 2024; Lee et al., 2024). They have also enhanced adaptive learning by analyzing performance patterns and proposing individualized learning pathways (Liu et al., 2024).

Nevertheless, their application is not without challenges. LLMs can reproduce biases present in their training data (Acerbi and Stubbersfield, 2023), generate inaccurate or unreliable content (Alqahtani et al., 2023), and require substantial technological infrastructure for deployment (Yan et al., 2024). Furthermore, their algorithmic "black-box" nature limits the traceability of their decisions (Xu et al., 2024), raising ethical concerns regarding privacy, equity, and transparency (Hadi et al., 2024; Meyer et al., 2024; Peláez-Sánchez et al., 2024). These issues have spurred the development of regulatory frameworks and bias mitigation strategies, although practical implementation remains uneven across regions and educational levels (Walter, 2024).

Rationale of the study

Despite the promising advances, the use of LLMs in educational settings presents substantial challenges that demand critical scrutiny. Recent studies underscore ethical and practical risks, including the generation of biased responses, the propagation of errors, and students' excessive dependence on these tools (Kasneci et al., 2023). Automated assessment via LLMs can compromise the validity of educational measurements (Caines et al., 2023), while their deployment in academic tasks has sparked debates over authorship and plagiarism (Meyer et al., 2023; Milano et al., 2023). Other works warn of the opacity of these

models, which hampers explainability and pedagogical control (Gan et al., 2023; Jeon and Lee, 2023). Moreover, uneven technical infrastructure among institutions creates an access gap that limits the effective use of these technologies. The need for clear pedagogical strategies and regulatory frameworks is widely acknowledged (Safranek et al., 2023), as is teacher training that enables the critical and contextualized integration of these tools (Roshanaei, 2024).

Against this backdrop, the present systematic review seeks to gather and analyze the existing literature on the implementation of advanced natural language processing models, such as those developed by OpenAI, Google, and Anthropic, in educational tasks, highlighting both their benefits and limitations. The primary aim is to examine how LLMs can optimize the generation of educational content, thereby allowing educators to devote greater attention to pedagogical interactions. This review also explores the opportunities and constraints identified in prior studies, with the goal of offering a comprehensive overview of the potential impact of these technologies and their applications within intelligent educational environments.

Research questions

The PICOS framework — Population, Intervention, Comparison, Outcomes, and Study design — facilitated the formulation of the research questions for this systematic review (Eriksen and Frandsen, 2018). By specifying these components, we defined the inclusion and exclusion criteria more precisely, strengthened the search strategy, and improved the consistency of study selection and critical appraisal.

To explore the factors that influence the adoption and effectiveness of LLMs in educational settings, this study synthesizes the available empirical evidence in the literature, addressing the following Research Questions (RQs):

- RQ1: How does the use of LLMs affect the time teachers devote to generating educational content and the perceived quality of that material compared with traditional methods?
- RQ2: How does the implementation of LLMs affect the automation of educational tasks and planning for teachers' pedagogical interactions in comparison with traditional practices?

Table 1 outlines the PICOS elements — Population (P), Intervention (I), Comparison (C), Outcomes (O), and Study design (S) — used to frame the two research questions (RQ1 and RQ2) concerning the integration of Large Language Models (LLMs) in education. RQ1 addresses reported changes in time devoted to content generation and perceived material quality. RQ2 examines reported support for planning-related tasks and task automation compared with traditional practices. Together, these elements provide a transparent structure for eligibility decisions and interpretation of findings.

Materials and methods

This systematic review was registered in the Open Science Framework (OSF) under the identifier <https://osf.io/v63nj/> and was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines

TABLE 1 PICOS elements for the research questions.

Element	RQ1	RQ2
P	Pre-tertiary and university teachers who face heavy workloads and constraints in generating personalized educational content.	Educators at institutions seeking to adapt to intelligent learning environments.
I	Use of LLMs as tools to generate educational content to alleviate teachers' workload.	Use of LLMs to automate administrative tasks and lesson preparation.
C	Traditional methods of generating educational content without AI-based LLMs.	Exclusive reliance on human resources for the same administrative tasks, without advanced technological tools.
O	Time devoted to generating educational content and perceived quality of instructional materials.	Extent of automation or task assistance and reported efficiency in pedagogical planning (e.g., time/steps devoted to planning and interaction).
S	Primary empirical studies and secondary evidence syntheses addressing RQ1.	Primary empirical studies and secondary evidence syntheses addressing RQ2.

(Page et al., 2021). The review was designed to examine the use of LLMs in educational settings.

To ensure a rigorous and transparent process in the selection and analysis of the included studies, specialized digital tools were employed at different stages of the review: (1) Rayyan Web (Rayyan Systems Inc.), an open-access platform designed to support systematic reviews, was used primarily during the title and abstract screening phase. This tool facilitates reviewer collaboration, labels studies by inclusion/exclusion criteria, detects conflicts, and allows for blinded decision-making, thereby improving traceability and reducing potential bias. (2) For data extraction and qualitative analysis, RevMan Web (Review Manager, developed by The Cochrane Collaboration) was employed. This platform enables researchers to structure study characteristics, graphically display results, and produce narrative syntheses. In the present study, RevMan Web was used exclusively to organize extracted data and systematize the presentation of findings.

To ensure both internal and external validity of the included studies, we applied standardized quality-assessment instruments appropriate to each study design.

- For non-randomized and quasi-experimental studies, the ROBINS-I tool developed by the Cochrane Collaboration was used to assess risk of bias across key domains, rating each as low, moderate, serious, or critical risk (Sterne et al., 2016). ROBINS-I evaluates risk of bias across seven key domains: (1) confounding, (2) selection of participants, (3) classification of interventions, (4) deviations from intended interventions, (5) missing outcome data, (6) measurement of outcomes, and (7) selection of reported results. Each domain is rated on a qualitative scale, low risk, moderate risk, high risk, or critical risk, providing a detailed appraisal consistent with the rigor required in systematic reviews.

- For qualitative, mixed-methods, and secondary systematic studies, the CASP checklist was applied to evaluate methodological rigor, clarity of aims, and coherence of findings (Critical Appraisal Skills Programme, 2022). Tailored versions of the CASP checklists were applied according to study design: (1) qualitative studies: The CASP Qualitative Checklist was used to assess clarity of the research aim, methodological appropriateness, researcher participant relationships, rigor of analysis, and relevance of findings. (2) To secondary systematic reviews: The CASP Systematic Review Checklist evaluated the explicit formulation of the research question, comprehensiveness of the search strategy, quality of included studies, and consistency of results and conclusions.

These guidelines enable a structured, standardized critical appraisal, fostering transparency and comparability across studies, particularly in reviews that employ mixed or narrative methodologies.

Eligibility criteria

Only records that simultaneously met all of the following criteria were eligible: (1) peer-reviewed journal articles or peer-reviewed conference proceedings; (2) written in English or Spanish and published within the predefined review period; (3) addressed the use of Large Language Models (LLMs) in educational settings with a clear teacher-facing component (e.g., instructional planning, content generation, feedback, assessment support, or related administrative/pedagogical tasks); (4) provided either primary empirical evidence (quantitative, qualitative, or mixed-methods) or secondary evidence syntheses (systematic reviews and narrative reviews) that directly synthesized evidence relevant to our research questions; and (5) reported outcomes or qualitative findings that could be mapped to at least one of our outcome domains (RQ1: efficiency/material quality; RQ2: automation/planning support). Importantly, eligibility did not require positive effects: studies reporting neutral or negative findings were eligible if they met the above criteria. We excluded grey literature (e.g., theses, technical reports, blogs), non-peer-reviewed opinion pieces without an explicit evidence-synthesis basis, and studies outside educational contexts. When multiple reports described the same underlying work, the most complete version was retained.

Information sources

To identify relevant literature, we searched three bibliographic sources that provide broad multidisciplinary coverage and strong representation of computer science and educational technology outlets: Scopus, ACM Digital Library, and Dimensions. Because our topic is framed around both pre-tertiary and university teachers, we acknowledge that we did not search education-specialist databases such as ERIC or PsycINFO, nor Web of Science education collections. This choice may underrepresent mainstream education research, particularly in K–12 contexts, and therefore limits the generalizability of our synthesis to “teachers” broadly. Full database search strategies are provided in Appendix A.

Methodology

To guarantee comprehensive and precise retrieval of literature on the educational applications of LLMs, we devised a search strategy that integrates free-text terms with controlled vocabulary. Initially, key phrases were compiled in two domains — education and artificial intelligence —. To reinforce consistency and minimize noise, descriptors drawn from the IEEE Thesaurus were incorporated, as this resource provides standardized terminology widely recognized in the fields of engineering and technology. The use of this vocabulary ensured both precision and uniformity in searching and classifying the scientific literature, facilitated the identification of relevant studies, and enabled meaningful comparisons across investigations. Moreover, employing standardized terms mitigated ambiguity and enhanced communication among researchers, fostering a clear and consistent understanding of concepts and technologies related to LLMs in education.

The search strings were organized around two conceptual clusters.

- The first cluster — focused on educational applications of LLMs — combined phrases such as: “Educational Transformation,” “Intelligent Learning Environments,” “AI-Assisted Tools,” “LLM-Based Tutoring Systems,” “Automated Content Generation,” and “Educational Outcomes.”
- The second cluster captured core artificial intelligence terminology, including: “Large Language Models,” “Natural Language Processing,” “Generative Pre-trained Transformer,” “Generative Adversarial Networks,” “Text Summarization,” and “Chatbots.”

By interweaving terms from both clusters with Boolean operators, the strategy maximized the retrieval of literature at the intersection of LLMs technology and educational practice. Because the query terms were aligned with our efficiency and planning-support focus, we acknowledge that relevant studies not framed using time/automation terminology may have been underretrieved; this potential selection bias is addressed in the Limitations section.

Figure 1 illustrates the query framework implemented in Scopus, the ACM Digital Library, and Dimensions to identify studies on the educational applications of LLMs. This framework guided the construction of platform-specific search equations — combining educational key phrases with AI-related terms using the AND operator — and informed the final selection based on inclusion criteria (peer

review, relevance to educational planning, and comparison of tools), ensuring broad and consistent coverage of the relevant literature.

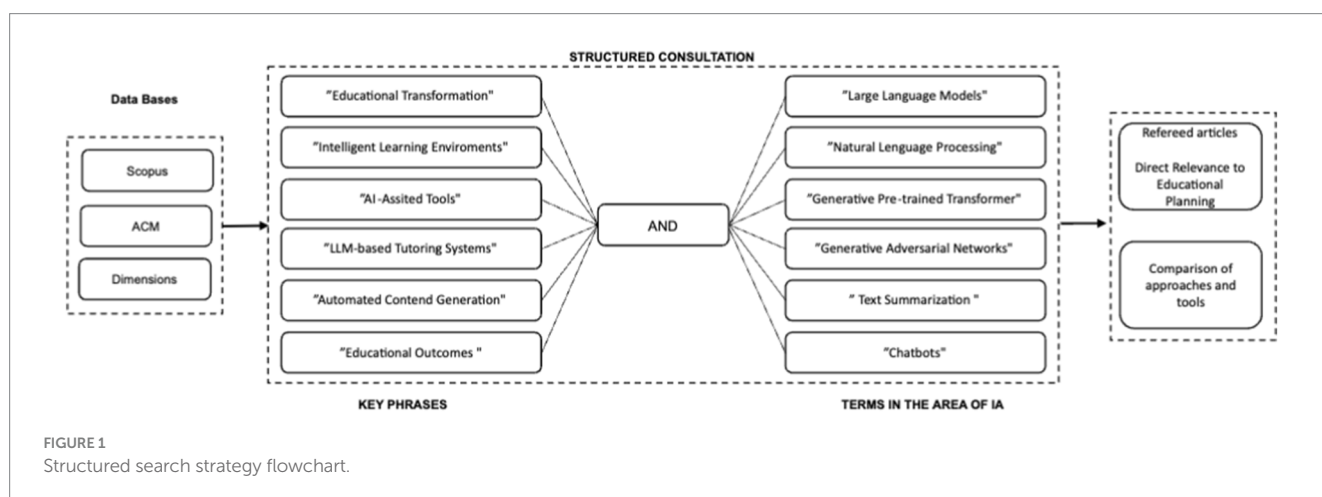
Study selection process

All records extracted from Scopus, ACM Digital Library, and Dimensions were imported into Rayyan, where, in an initial phase, the platform automatically removed duplicates. We then adopted a two-stage selection procedure conducted independently by three reviewers. (1) In the screening stage, the three reviewers independently examined the titles and abstracts of the remaining studies, discarding — without further assessment — any that did not meet the inclusion criteria. When discrepancies arose — e.g., if one or more reviewers labeled a study as “include” and others as “exclude,” the case was discussed until consensus was reached; if consensus could not be achieved, a fourth reviewer was brought in to make the final decision. (2) The full texts of the preselected articles were then evaluated, again independently, against the predefined eligibility criteria. Any disagreement was resolved by consensus between the principal reviewers or, failing that, through the intervention of the fourth reviewer. Given the database scope, we interpreted findings with caution regarding coverage of education-focused venues.

Data extraction

To ensure exhaustive and reliable information retrieval, all included articles underwent a three-step double-review process: (1) the authors independently assessed the full text of each study, recording in a predefined template the following variables: author(s), year of publication, country or educational context, type of LLMs employed, methodological design, educational level, and characteristics of the participant population; (2) the research objectives and the quantitative and qualitative results related to the research questions were identified; (3) the authors compared their records and resolved any discrepancies through discussion until consensus was reached.

Where counts are reported (e.g., n/N), N refers to the denominator specified in each statement. Unless otherwise noted, frequency tallies are based on primary empirical studies ($N = 13$), while secondary evidence syntheses ($N = 3$) are described separately to avoid conflating empirical outcomes with review-level narrative conclusions.



Data items

- To answer our research questions, we grouped the outcomes into two clearly defined domains:
- Efficiency and Perceived Material Quality (RQ1): time devoted to content generation and perceived quality of instructional materials.
- Automation and Planning Support (RQ2): reported extent of automation or task assistance in teaching and indicators of planning efficiency.

Subsequently, we extracted from each study quantitative and qualitative data compatible with these domains, including time measures (minutes or hours), perceived quality ratings, reported automation indicators (e.g., self-reported proportions where available), planning efficiency indicators (e.g., steps reduced or time saved), and participants' comments and thematic findings.

Unless otherwise stated, references to improvements in material quality, clarity, or usefulness reflect perceived evaluations — teacher- or learner-rated scales — rather than external objective assessment.

In addition to these outcomes, we cataloged and defined all contextual variables required to interpret the findings and guarantee comparability:

- Author(s) and year of publication, to establish the chronological position of each study.
- Country or educational setting (secondary or higher education).
- Type of LLM employed (e.g., GPT-3, BERT, T5).
- Methodological design (experimental, quasi-experimental, survey, case study, or observational).
- Educational level and sample characteristics, including size, age range, and discipline.
- Stated objectives of the study, as reported in each article.

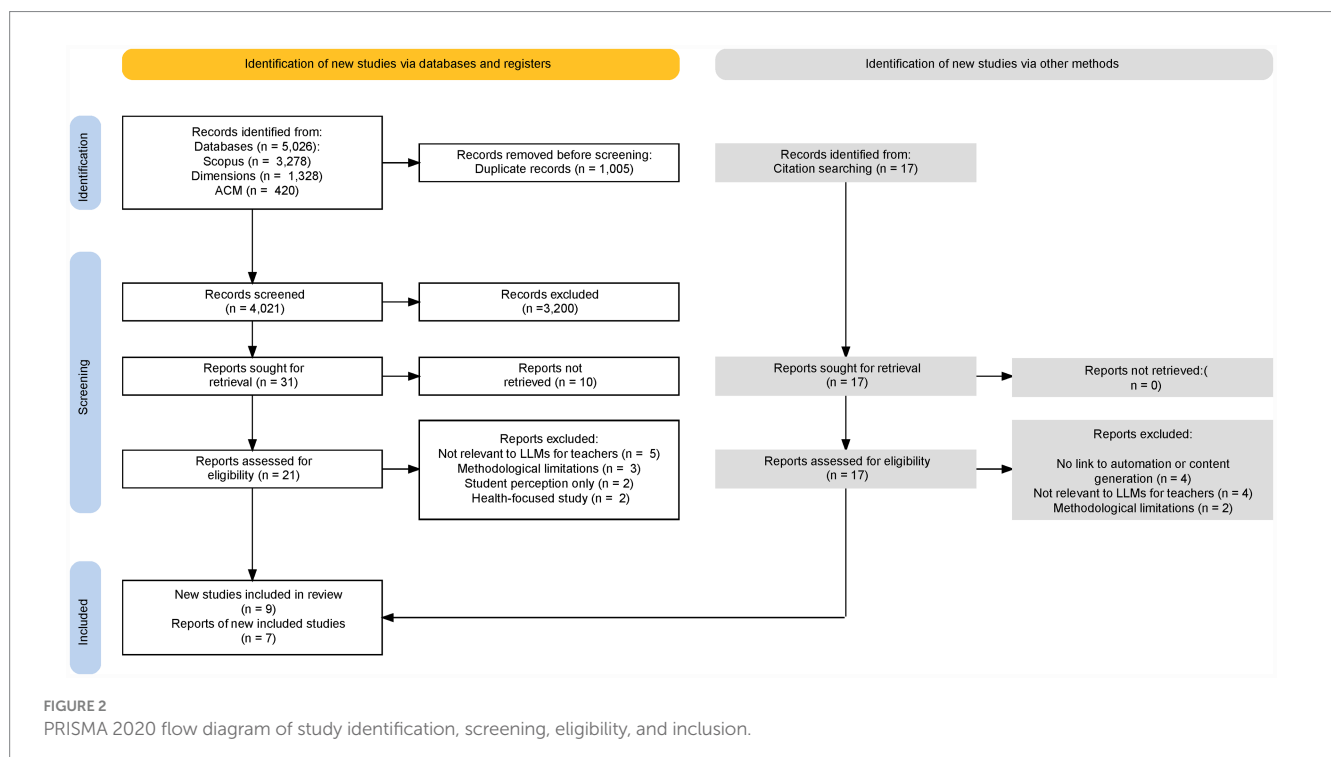
Risk-of-bias assessment

To safeguard the robustness of the findings, every included study underwent a rigorous risk-of-bias evaluation. For non-randomized studies, the ROBINS-I tool was applied, with two reviewers independently rating seven bias domains: (1) confounding variables, (2) participant selection, (3) intervention classification, (4) deviations from the intended intervention, (5) incomplete outcome data, (6) blinding of outcome assessment, and (7) selective reporting. Each domain was graded as low, moderate, or high risk, and discrepancies between reviewers were resolved first through discussion and, if necessary, through the involvement of a third reviewer. For qualitative or mixed-methods investigations, the CASP checklist for qualitative and mixed design studies was used, examining the clarity of objectives, methodological relevance, design appropriateness, recruitment representativeness, data quality, researcher-participant relationships, ethical considerations, analytical rigor, clarity of results, and transferability of conclusions. Similarly, for secondary systematic reviews, the CASP tool for reviews evaluated the validity of the design, methodological soundness, presentation of results, and applicability of the recommendations.

Results

Selection of literature

Figure 2 illustrates the bibliographic selection process followed in this systematic review. A total of 5,026 records were retrieved from Scopus ($n = 3,278$), Dimensions ($n = 1,328$), and the ACM Digital Library ($n = 420$). After removing 1,005 duplicates, a total of 4,021 titles and abstracts remained for screening, of which 3,200 were excluded for not meeting the eligibility criteria. Thirty-one full-text reports were sought; 10 could not be obtained, and 21 were fully assessed. Of these, 12 were excluded (5 not pertinent to LLMs in



teaching; 3 with methodological limitations; 2 focused exclusively on student perception; 2 centered on health-related issues), leaving 9 studies to advance to the quantitative/qualitative synthesis. An additional 17 reports were located through citation tracking; all were retrieved and evaluated, but 10 were excluded (4 without links to automation or content generation; 4 irrelevant to LLMs in teaching; 2 with methodological problems), so 7 were ultimately included. In total, 16 studies met the inclusion criteria and were integrated into the review.

Characteristics of the included studies

Sixteen records published between 2023 and 2025 were included: 13 primary empirical studies — qualitative, quantitative, quasi-experimental, and mixed-methods designs — and three secondary evidence syntheses — one systematic review and two narrative/theoretical reviews. Most evidence originates from higher-education settings and short-term deployments.

Across the primary studies, outcomes and qualitative findings were mapped to RQ1 (efficiency/material quality) and/or RQ2 (automation/planning support), but measurement approaches were heterogeneous and frequently perception-based. Secondary evidence syntheses were used to contextualize patterns and gaps rather than as equivalent empirical effect estimates.

Table 2 provides a condensed overview of each record. The complete extraction table (including objectives, detailed results, and RQ1/RQ2 outcome fields) is provided in Supplementary Table S1.

Methodological design of the studies

Figure 3 summarizes the methodological designs of the 16 included records, reflecting an emerging and heterogeneous evidence base. Three qualitative studies (Al-Mughairi and Bhaskar, 2024; Isiaku et al., 2024; Sajadi et al., 2024) used interviews or longitudinal case-study approaches to explore experiences and perceived usefulness. Four studies used quantitative or quasi-experimental/comparative designs (Abreu et al., 2024; Gasaymeh and AlMohtadi, 2024; Ma et al., 2023; Winder et al., 2024), typically comparing outcomes with and without LLM support or validating generated outputs. Six studies employed mixed-methods designs combining surveys, content analysis, intervention data, and qualitative feedback (Ivanović, 2023; Jyothy et al., 2024; Lee and Song, 2024; Magalhães Araujo and Cruz-Correia, 2024; Song et al., 2024; Veras et al., 2024). Finally, three secondary evidence syntheses (Albadarin et al., 2024; Wang et al., 2024; Shahzad et al., 2025) summarized empirical literature to describe applications, risks, and research trends. This distribution indicates a field still in consolidation, where exploratory user studies and early comparative evaluations coexist with a smaller number of evidence syntheses.

Educational levels addressed by the studies

Table 3 shows that most of the included studies are situated in higher education, encompassing undergraduate and graduate programs as well as teacher-training initiatives. Two studies did not specify the educational level, while another one is distributed across mixed settings (basic, higher, and informal), informal medical education, and university level

programming courses. This diversity indicates that LLMs have been examined both in formal higher-education contexts and in more flexible, informal learning environments, highlighting their versatility across academic and professional settings.

Sample size and study objectives

The participant samples varied widely: from a content-analysis study with no human participants (Abreu et al., 2024) to surveys involving 81 students and 11 teachers (Lee and Song, 2024), as well as a review encompassing 50 separate studies (Albadarin et al., 2024). Among the empirical investigations with quantitative samples, the median number of participants was roughly 70, whereas the qualitative studies ranged from 30 to 100 participants, including students, instructors, and research assistants. Three main research lines predominated:

- Several studies focused on assessing teachers' and students' perceptions of the usefulness of LLMs in instruction (Song et al., 2024; Lee and Song, 2024).
- Comparative investigations measured efficiency indicators and the readability of materials generated with and without LLM support (Abreu et al., 2024; Gasaymeh and AlMohtadi, 2024).
- Targeted research explored the application of these tools in specific contexts — academic research, medical education, and teacher training — examining their capacity to enhance planning processes and pedagogical feedback (Al-Mughairi and Bhaskar, 2024; Isiaku et al., 2024; Winder et al., 2024).

This heterogeneity in both sample size and purpose highlights an emerging field in which researchers are examining user experiences as well as the performance metrics of LLMs in educational environments.

Educational content generated

Figure 4 shows that 14 of the studies employed LLMs to generate a wide range of educational content: from personalized answers and practice exercises (Song et al., 2024) to the drafting of essays, outlines, and structured study guides (Albadarin et al., 2024; Shahzad et al., 2025). Several investigations explored the creation of assessment questions, rubrics, and classroom activities, with examples drawn from programming and computer science courses (Ma et al., 2023; Sajadi et al., 2024; Winder et al., 2024). In the health domain, one study rewrote medical materials tailored for oncology patients (Abreu et al., 2024), while others produced clinical narratives and pedagogical simulations (Al-Mughairi and Bhaskar, 2024). LLMs were also used for lesson planning, slide presentations, and proposals for interactive tasks (Gasaymeh and AlMohtadi, 2024; Isiaku et al., 2024; Jyothy et al., 2024; Veras et al., 2024), as well as for article synthesis and the preparation of executive summaries (Ivanović, 2023). Only a single study did not report generating didactic materials (Wang et al., 2024). Taken together, these findings emphasize that the principal instructional application of LLMs lies in the production of diversified teaching resources that support educators in planning and students in self-directed learning.

TABLE 2 Condensed basic characteristics of the included studies (full extraction in [Supplementary Table S1](#)).

Reference	Country/Context	Educational level	Study type/Design	LLM	Key outcomes: RQ1/RQ2
Ivanović (2023)	Montenegro	University	Mixed study (comparative analysis + qualitative analysis)	ChatGPT (version not specified)	<ul style="list-style-type: none"> Content: No educational materials were generated, only automated feedback. Automation: Yes, trial evaluation and detailed feedback generation was automated. Planning: It does not address planning or teacher-student educational interaction.
Albadarin et al. (2024)	Global/General education	Mixed	Systematic review of empirical studies	ChatGPT (version not specified)	<ul style="list-style-type: none"> Content: Yes, such as writing texts, outlines, ideas, evaluations and assignments. Automation: Yes, automation of feedback, improvement suggestions, quiz and homework generation. Planning: Support for instructional planning and instructional design is mentioned.
Lee and Song (2024)	Finland	Higher education	Qualitative and quantitative comparative perception study	GPT – 3.5	<ul style="list-style-type: none"> Content: Yes, automated generation of conceptual explanations using LLM. Automation: Partial: Automation of the generation of explanations, but not of broader pedagogical processes. Planning: It does not address planning or teacher-student interaction directly.
Abreu et al. (2024)	USA	Informal medical education	Quantitative study of generated content and comparative analysis	ChatGPT 4.0	<ul style="list-style-type: none"> Content: Yes, ChatGPT rewrote patient education content about cancer. Automation: Yes, automation of rewriting complex content to make it more accessible to the general. Planning: It does not address educational planning or interaction between teachers and students.
Isiaku et al. (2024)	Northern Cyprus	Postgraduate / Higher Education	Qualitative study with semi-structured interviews.	ChatGPT (version not specified)	<ul style="list-style-type: none"> Content: Yes, generation of academic content such as essays, articles, emails and research proposals. Automation: Yes, tasks such as writing, proofreading, idea generation and information synthesis are automated. Planning: Partially: learning is facilitated, but formal educational planning is not addressed.
Winder et al. (2024)	Switzerland	Higher education/teacher training	Comparative quantitative study (with and without AI)	GPT-4, Ollama, HuggingFace	<ul style="list-style-type: none"> Content: Yes, creation of complete educational modules, questions, explanatory texts and visuals. Automation: Yes, automation of texts, images, evaluations and concepts in the modules. Planning: Yes, support for instructional planning and pedagogical alignment between objectives and content.
Veras et al. (2024)	Australia	University Health sciences	Mixed methods randomized crossover trial (quantitative + qualitative)	ChatGPT (version not specified)	<ul style="list-style-type: none"> Content: Yes, students used ChatGPT to generate summaries, explanations, answers to questions. Automation: Partial: ChatGPT assisted in learning tasks, but not teaching or administrative tasks. Planning: It does not address planning or teacher pedagogical interaction.

(Continued)

TABLE 2 (Continued)

Reference	Country/Context	Educational level	Study type/Design	LLM	Key outcomes: RQ1/RQ2
Gasaymeh and AlMohtadi (2024)	Jordan	University	Quasi-experimental study	ChatGPT (version not specified)	<ul style="list-style-type: none"> Content: Yes, ChatGPT supported the elaboration of explanations, problem solving and homework preparation. Automation: Yes, ChatGPT helped to resolve doubts, reinforce topics and support learning outside the classroom. Planning: Promotes classroom interaction and active learning, although no teacher planning was evaluated.
Sajadi et al. (2024)	USA	University	Qualitative longitudinal case study	ChatGPT 4.0	<ul style="list-style-type: none"> Content: Yes, generation of written feedback from peer and self-assessment data. Automation: Yes, the synthesis and writing of individualized feedback reports was automated. Planning: Teachers designed prompts, reviewed and edited results; there was clear pedagogical alignment.
Al-Mughairi and Bhaskar (2024)	Oman	University	Qualitative study with semi-structured interviews.	ChatGPT (version not specified)	<ul style="list-style-type: none"> Content: Yes, generation of evaluations, class guides, support materials and thematic explanations. Automation: Yes, processes of writing, preparation of materials and support in educational tasks were automated. Planning: Yes, ChatGPT was used to plan lessons, explain concepts and adapt content to student.
Magalhães Araujo and Cruz-Correia (2024)	Portugal	Postgraduate (Master's Degrees)	Mixed study (quantitative survey + qualitative analysis)	ChatGPT (version not specified)	<ul style="list-style-type: none"> Content: Yes, used to create academic content, exams, clinical guides, explanations and exercises. Automation: Yes, support for the generation of evaluations, educational material, class preparation and simulations. Planning: Yes, teaching proposals are included for use in planning, guided interaction and simulated clinical...
Song et al. (2024)	USA	University	Intervention	ChatGPT-3.5, ChatGPT-4 Turbo, ChatGe-V1/V2	<ul style="list-style-type: none"> Content: Yes, especially in the development of answers, exercises and customized content. Automation: Yes, responses, feedback and part of the programming tasks were automated. Planning: Yes, ChatGPT helped in post-reading interaction and teacher-led activities.
Jyothy et al. (2024)	China	University	Mixed methods study (quantitative and qualitative)	ChatGPT-3.5, ChatGPT-4.0	<ul style="list-style-type: none"> Content: Yes, used to generate essays, presentations, guides, outlines and educational materials. Automation: Yes, automation of summaries, corrections, translations, outlines and content writing. Planning: Yes, used to create questions, structure lessons, write materials and plan lessons.
Ma et al. (2023)	China	University	Quasi-experimental study + expert validation	ChatGPT / HypoCompass (Own LLM)	<ul style="list-style-type: none"> Content: Yes, the LLM generated explanations, simulated bugs, code feedback and debugging practices. Automation: Yes, automation of material generation, error simulation, feedback and didactic support. Planning: Yes, it is structured around an instructional model with learning objectives.

(Continued)

TABLE 2 (Continued)

Reference	Country/Context	Educational level	Study type/Design	LLM	Key outcomes: RQ1/RQ2
Wang et al. (2024)	Global	Not specified	Narrative review	ChatGPT-4.0, Claude, LLaMA, PALM.	<ul style="list-style-type: none"> Content: Yes, generation of educational materials, exams, summaries, explanations, guides and questionnaires. Automation: Yes, automation of tasks such as feedback, exercise review, activity generation and automatic responses. Planning: Yes, its applications for lesson planning, structuring content and supporting personalized teaching are discussed.
Shahzad et al. (2025)	Global	Not specified	Narrative / theoretical review	ChatGPT-3.5, ChatGPT-4.0	<ul style="list-style-type: none"> Content: Yes, multiple applications are mentioned to generate explanations, exercises, guides, assignments. Automation: Yes, it describes the automation of feedback, question generation, task review and writing. Planning: Yes, its use as a support for instructional planning and personalization of learning is...

Pedagogical planning and interaction

Figure 5 shows that 14 of the 16 studies assessed the support of LLMs for educational content generation, task automation, and pedagogical planning/interaction (Abreu et al., 2024; Al-Mughairi and Bhaskar, 2024; Albadarin et al., 2024; Gasaymeh and AlMohtadi, 2024; Isiaku et al., 2024; Ivanović, 2023; Jyothy et al., 2024; Magalhães Araujo and Cruz-Correia, 2024; Sajadi et al., 2024; Shahzad et al., 2025; Song et al., 2024; Veras et al., 2024; Wang et al., 2024; Winder et al., 2024). Reported efficiency or ‘automation’ gains were operationalized using heterogeneous measures (e.g., minutes or hours saved, steps eliminated, or self-reported productivity/satisfaction). Therefore, we synthesize these outcomes narratively and treat study-specific numeric estimates as illustrative rather than pooled effect sizes.

Limitations of LLMs

Table 4 synthesizes the principal constraints reported across the reviewed studies. Importantly, these limitations should be interpreted as characteristics of the specific models, interfaces, and institutional configurations evaluated in the 2023–early-2025 literature, rather than stable or “inherent” properties of LLMs as a class. Reported constraints included (1) limited context windows and challenges integrating multiple sources for complex synthesis tasks, (2) knowledge cutoffs that can yield outdated responses for recent content, (3) output risks such as hallucinations and bias that require verification and governance, and (4) implementation and platform constraints — prompt-only workflows, limited LMS integration, and lack of reusable templates — that are largely product- and workflow-design limitations. Given the rapid pace of model and tooling development, these constraints are best treated as time-bounded and implementation-dependent, and conclusions should avoid projecting current limitations as permanent features of “LLMs in education.”

Summary of the functional and contextual limitations of LLMs reported in the reviewed studies, with emphasis on their applicability in teaching practice.

Design limitations for teaching use

Figure 6 shows the principal design constraints that condition teachers’ adoption of LLM-based tools: (1) reliance on textual prompts requires a high level of digital literacy and advanced language skills, which hampers seamless use in contexts with limited technological training (Al-Mughairi and Bhaskar, 2024; Isiaku et al., 2024; Ivanović, 2023; Jyothy et al., 2024; Lee and Song, 2024; Ma et al., 2023; Sajadi et al., 2024; Song et al., 2024). (2) The absence of integrated graphical interfaces within learning management systems forces users to switch between windows or applications, interrupting teachers’ workflow and reducing efficiency (Magalhães Araujo and Cruz-Correia, 2024; Veras et al., 2024; Winder et al., 2024). (3) The lack of automatic personalization — such as templates adapted to different grade levels or subject areas — limits the relevance of generated content and necessitates frequent manual adjustments (Abreu et al., 2024; Gasaymeh and AlMohtadi, 2024; Shahzad et al., 2025; Wang et al.,

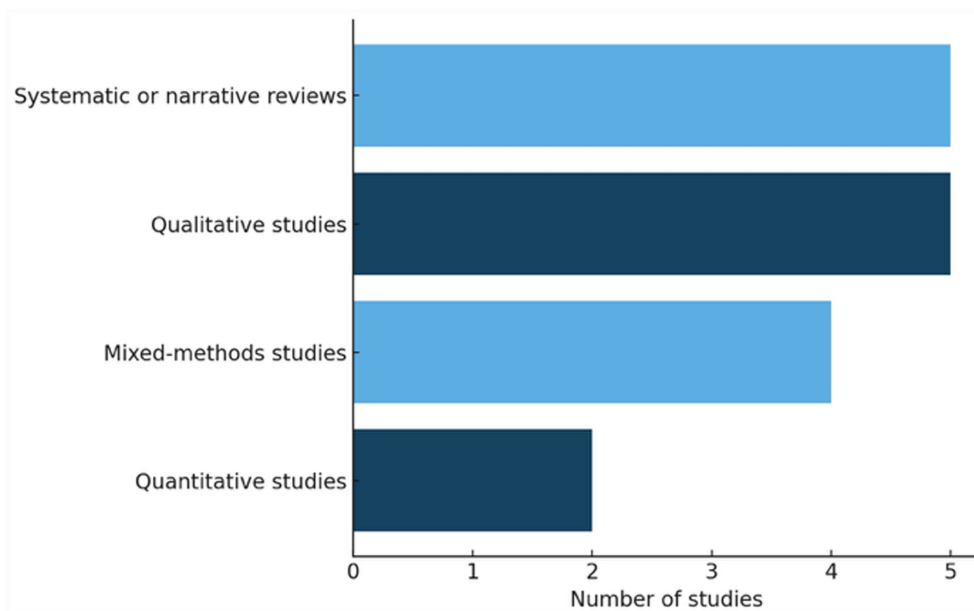


FIGURE 3 Distribution of methodological designs among the included studies.

TABLE 3 Studies by educational level.

Educational level	Studies
Higher education (undergraduate, graduate, teacher training, and specialized university courses)	Abreu et al. (2024), Al-Mughairi and Bhaskar (2024), Gasaymeh and AlMohtadi (2024), Isiaku et al. (2024), Ivanović (2023), Jyothy et al. (2024), Lee and Song (2024), Ma et al. (2023), Magalhães Araujo and Cruz-Correia (2024), Sajadi et al. (2024), Song et al. (2024), Veras et al. (2024), and Winder et al. (2024)
Mixed (basic, higher, and informal)	Albadarin et al. (2024)
Not specified	Shahzad et al. (2025) and Wang et al. (2024)

2024). (4) The absence of real-time feedback prevents instructors from correcting or refining outputs interactively during planning, often leading to time-consuming iterations (Albadarin et al., 2024). These limitations underscore the need to enhance the user experience — through clear graphical interfaces, intuitive visual panels, and streamlined workflows — and to improve both the integration and adaptability of LLMs in order to maximize their utility in everyday teaching practice.

Output limitations

Figure 7 summarizes the five main limitations identified in the outputs generated by LLMs across the 16 studies analyzed: (1) lack of reproducibility when even minimal changes are made to prompts, reported in 10 of 16 studies (Al-Mughairi and Bhaskar, 2024; Ivanović, 2023; Jyothy et al., 2024; Lee and Song, 2024; Ma et al.,

2023; Sajadi et al., 2024; Song et al., 2024; Veras et al., 2024; Wang et al., 2024; Winder et al., 2024), which hinders the replication of findings; (2) insufficient content coverage, detected in 14 studies (Abreu et al., 2024; Albadarin et al., 2024; Al-Mughairi and Bhaskar, 2024; Gasaymeh and AlMohtadi, 2024; Isiaku et al., 2024; Ivanović, 2023; Jyothy et al., 2024; Lee and Song, 2024; Magalhães Araujo and Cruz-Correia, 2024; Sajadi et al., 2024; Shahzad et al., 2025; Song et al., 2024; Veras et al., 2024; Winder et al., 2024), indicating that models omit relevant aspects when instructions are not sufficiently precise; (3) factual errors or “hallucinations,” documented in 8 studies (Albadarin et al., 2024; Gasaymeh and AlMohtadi, 2024; Isiaku et al., 2024; Jyothy et al., 2024; Ma et al., 2023; Magalhães Araujo and Cruz-Correia, 2024; Shahzad et al., 2025; Wang et al., 2024) (8 of 16 studies), undermining the reliability of the generated information; (4) security and privacy concerns when sensitive data are included in prompts, raised in 7 studies (Abreu et al., 2024; Al-Mughairi and Bhaskar, 2024; Ivanović, 2023; Lee and Song, 2024; Ma et al., 2023; Shahzad et al., 2025; Winder et al., 2024) (7 of 16 studies); and (5) dependence on input quality, noted in 5 studies (Gasaymeh and AlMohtadi, 2024; Isiaku et al., 2024; Lee and Song, 2024; Ma et al., 2023; Song et al., 2024) (5 of 16 studies), showing that the models’ effectiveness drops sharply when prompts are poorly formulated. Recognizing these limitations is essential for guiding the design of future research and for improving best practices in the educational use of LLMs.

Non-reproducibility

Several reviewed studies ($n = 10$) reported challenges reproducing LLM outputs across repeated uses (Al-Mughairi and Bhaskar, 2024; Ivanović, 2023; Jyothy et al., 2024; Lee and Song, 2024; Ma et al., 2023; Sajadi et al., 2024; Song et al., 2024; Veras et al., 2024; Wang et al., 2024; Winder et al., 2024). However, in the reviewed evidence base,

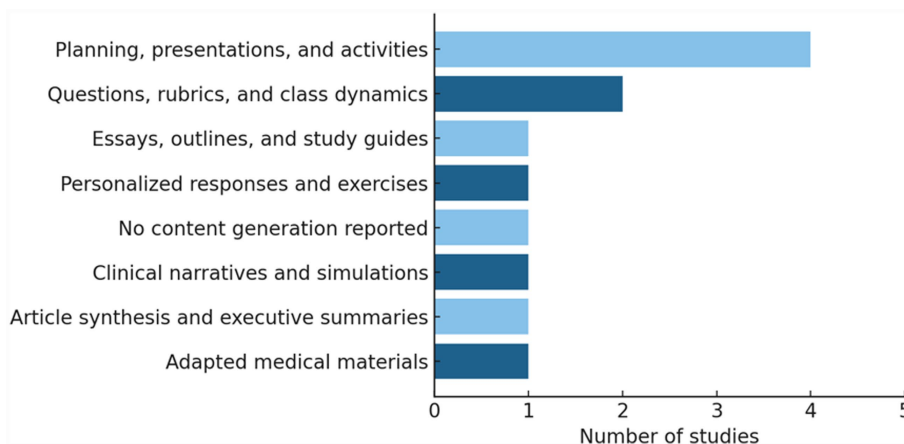


FIGURE 4 Types of educational content generated with LLMs across the studies.

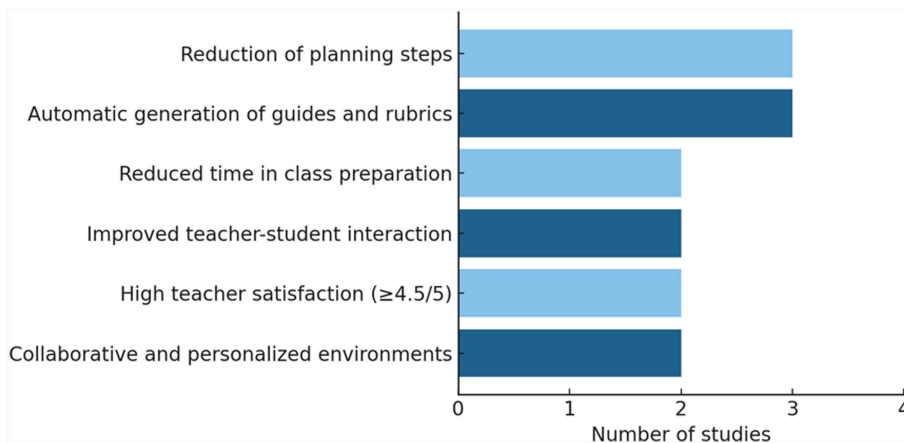


FIGURE 5 Support provided by large language models for pedagogical planning and teacher interaction.

TABLE 4 Limitations of LLMs in educational contexts.

Identified limitation	Description
Limited context window	Makes it difficult to handle lengthy documents or to integrate multiple sources.
Lack of data updates	Models lack recent information because of their cutoff date.
Model opacity (“black box”)	It is not understood how the answers are generated, which limits their validation.
Biases in the training corpus	Reproduce stereotypes and cultural biases present in the source data.
High computational cost	Require advanced infrastructure, which poses a challenge in low-resource settings.

non-reproducibility often reflects a combination of model stochasticity and tooling/reporting practices — such as using consumer chat interfaces, not pinning model versions, and incomplete reporting of prompts, temperature, and other generation parameters — rather than

an intrinsic barrier that is immutable. We therefore frame this issue as a limitation of the models, platforms, and reporting practices used in the reviewed studies. Future work should adopt reproducibility-oriented reporting checklists (model/provider, version, date, system prompts, full user prompts, and generation settings) and, where feasible, use version-pinned APIs and archived prompt sets to enable replication.

Lack of exhaustiveness

Fourteen of the 16 studies (Abreu et al., 2024; Al-Mughairi and Bhaskar, 2024; Albadarin et al., 2024; Gasaymeh and AlMohtadi, 2024; Isiaku et al., 2024; Ivanović, 2023; Jyothy et al., 2024; Lee and Song, 2024; Magalhães Araujo and Cruz-Correia, 2024; Sajadi et al., 2024; Shahzad et al., 2025; Song et al., 2024; Veras et al., 2024; Winder et al., 2024) warned that LLMs omitted relevant information when instructions were insufficiently precise, compromising the exhaustiveness of the results. These omissions ranged from key contextual data — such as specific pedagogical objectives — to essential methodological details, potentially biasing the interpretation

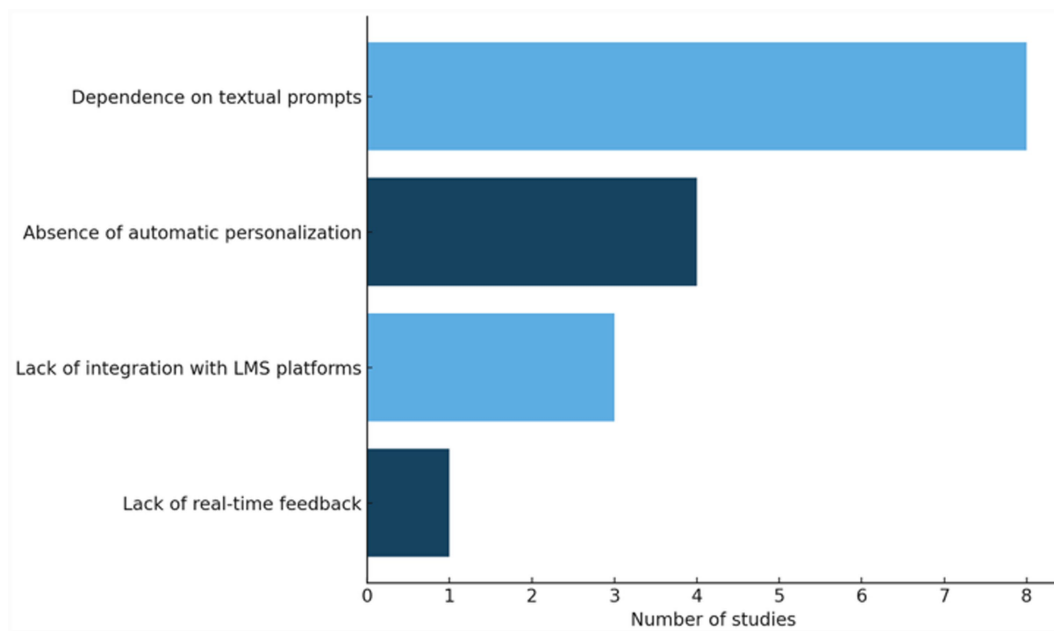


FIGURE 6 Main design limitations affecting the adoption of LLM-based tools by teachers.

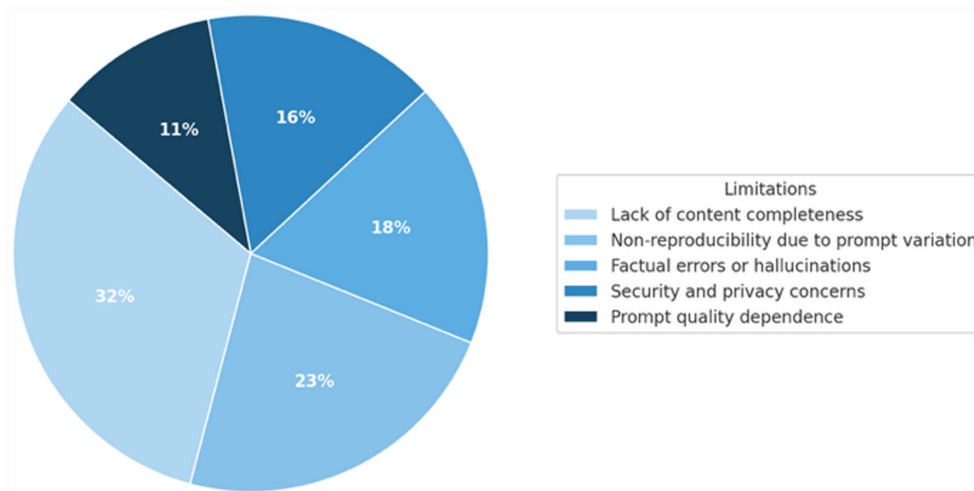


FIGURE 7 Main limitations identified in large language model outputs across the reviewed studies.

of the generated materials. Without thorough manual validation or review, automated content risks overlooking fundamental concepts or critical examples. To improve coverage, more structured “prompt engineering” protocols and the use of checklists are recommended to ensure that all intended pedagogical elements are included.

Risk of bias in the included studies

ROBINS-I was applied to evaluate the two quasi-experimental studies (Gasaymeh and AlMohtadi, 2024; Winder et al., 2024); CASP checklists were used for the qualitative and mixed-methods studies in

11 papers (Abreu et al., 2024; Al-Mughairi and Bhaskar, 2024; Ivanović, 2023; Isiaku et al., 2024; Jyothy et al., 2024; Lee and Song, 2024; Ma et al., 2023; Magalhães Araujo and Cruz-Correia, 2024; Sajadi et al., 2024; Song et al., 2024; Veras et al., 2024), and the CASP checklist for reviews was employed in three studies (Albadarin et al., 2024; Shahzad et al., 2025; Wang et al., 2024).

The aggregated risk profiles by domain are presented below, followed by the traffic light matrices for each individual study.

Figure 8 illustrates the bias levels identified with ROBINS-I in the quasi-experimental studies (Gasaymeh and AlMohtadi, 2024; Winder et al., 2024). For five of the seven domains — control of confounding variables, selection of participants, deviations from the

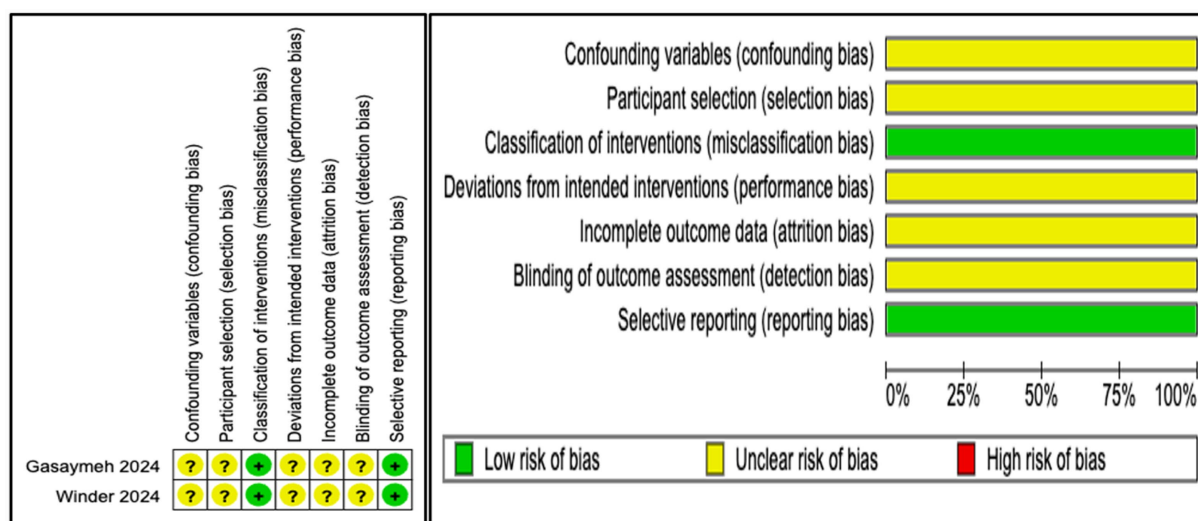


FIGURE 8 ROBINS-I risk-of-bias assessment for quasi-experimental studies. (a) Aggregated domain ratings. (b) Traffic-light matrix showing risk levels for each bias domain.

intended intervention, incomplete outcome data, and blinding of outcome assessment — the risk was classified as unclear (no information) in both studies, whereas only the domains of classification of interventions and selective outcome reporting showed low risk. No domain was rated as high risk, but the prevalence of uncertainties underscores the need to strengthen bias control at key stages of quasi-experimental study design. Both quasi-experimental studies showed ‘unclear’ risk of bias in five of the seven ROBINS-I domains, underscoring substantial uncertainty in this evidence.

As detailed in Figure 9, 11 qualitative and mixed-methods studies were evaluated using the CASP checklist, and the domains of clarity of objectives, data quality, and contribution/transferability all received unanimously low-risk ratings. However, uncertainties were recorded for methodological relevance (3 studies), analytical rigor (4), and the researcher-participant relationship (2), along with a high risk in the domain of representativeness of recruitment. These results suggest that improving sampling strategies and enhancing researcher reflexivity are essential to increase the robustness of qualitative findings.

As shown in Figure 10, the three systematic reviews assessed with the CASP checklist exhibited low risk in the domains of design, methodological robustness, and applicability. The only element with uncertain risk was the presentation of results, which was rated as uncertain in one of the studies. This consistency reflects rigorous methodological handling, although clearer reporting of findings is recommended to enhance transparency.

Taken together, the risk-of-bias analyses reveal both strengths and areas for improvement. With ROBINS-I, uncertainties predominate in five critical domains despite the low-risk observed in intervention classification and selective reporting. With CASP for qualitative/mixed-methods, clarity of objectives and data quality were solid,

although concerns emerged regarding recruitment and analytical rigor. Finally, low-risk was recorded for design and applicability, with only one results domain remaining uncertain, underscoring the need to improve informational transparency.

Summary of key findings and research gaps

To provide a concise overview of what the current evidence suggests—while respecting heterogeneity and risk-of-bias—Table 5 summarizes the main trends, measurement approaches, and remaining gaps. Findings should be interpreted as indicative patterns rather than definitive effect estimates.

Discussion

This systematic review synthesized 16 studies on the use of LLMs in educational contexts, focusing on two research questions: efficiency and perceived material quality (RQ1) and task automation/assistance and planning support (RQ2). Notably, the corpus is skewed toward higher-education contexts and frequently involves small or short-term samples; moreover, risk-of-bias appraisal indicated several domains rated as unclear, limiting confidence in the magnitude and generality of reported effects. Overall, the included literature suggests that LLMs can support instructional workflows and that teachers and learners often report perceived benefits — clarity, convenience, or usefulness of generated resources. However, the evidence base is heterogeneous in study designs and outcome definitions, frequently relies on self-reported measures, and includes secondary evidence syntheses; therefore, findings should be interpreted cautiously and not as pooled effect estimates.

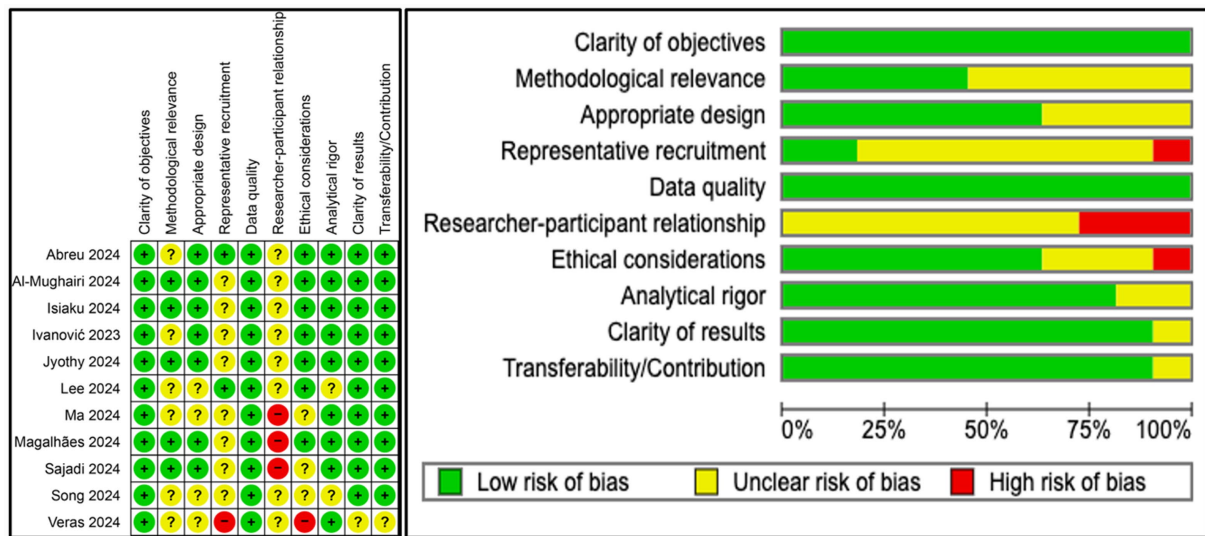


FIGURE 9 CASP appraisal for qualitative and mixed-methods studies. (a) Aggregated ratings across assessment domains. (b) Traffic-light matrix showing risk levels for each checklist item.

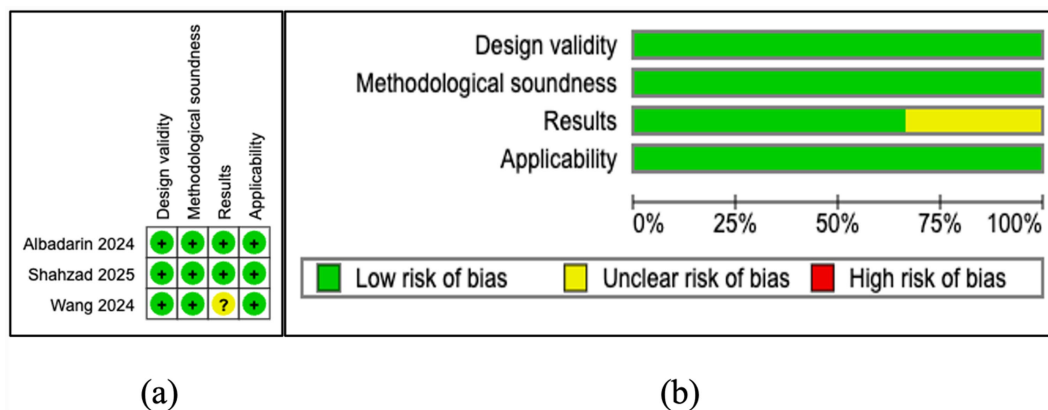


FIGURE 10 CASP appraisal for systematic reviews. (a) Aggregated ratings across assessment domains. (b) Traffic-light matrix showing risk levels for each checklist item.

Pedagogical framing and interpretation

To strengthen the pedagogical framing, we interpret the reviewed evidence through the lens of instructional planning as an iterative design-and-feedback cycle and through teacher workload management. From this perspective, LLMs most consistently support the development and revision phases (e.g., drafting lesson materials, generating variations, and producing first-pass feedback), while pedagogically high-stakes decisions (alignment with learning objectives, assessment validity, differentiation, and equity) remain the responsibility of educators. Accordingly, the most defensible implementation model across the reviewed studies is human-in-the-loop:

teachers specify pedagogical intent, use LLMs for initial drafts or ideas, and then verify, adapt, and contextualize outputs before classroom use (Kasneci et al., 2023; Jeon and Lee, 2023).

Output variability in response to minimal prompt modifications remains a major challenge. Ten studies ($n = 10$) reported substantial differences in outputs that were semantically similar but diverged in critical pedagogical content (Al-Mughairi and Bhaskar, 2024; Ivanović, 2023; Jyothy et al., 2024; Lee and Song, 2024; Ma et al., 2023; Sajadi et al., 2024; Song et al., 2024; Veras et al., 2024; Wang et al., 2024; Winder et al., 2024). This instability prevents the establishment of replicable protocols: an instructor who follows the same set of instructions in two sessions may receive different materials, complicating the comparison of interventions and the validation of best practices.

TABLE 5 Summary of key findings, measurement approaches, and research gaps across outcome domains.

Domain (RQ)	What studies tend to report	How it is measured	Key limitations in evidence	Priority research gaps/next steps
RQ1: Time and perceived material quality	Reports of time savings and/or reduced preparation burden, alongside perceived improvements in material clarity or alignment, particularly when prompts are well-specified.	Objective time logs (minutes/h) in a subset; otherwise self-reported time saved or perceived efficiency; perceived quality ratings (teacher/student scales) and qualitative feedback.	Small samples; heterogeneity of tasks (lesson plans, feedback, quizzes); varying LLM versions and interfaces; reliance on subjective measures; limited controlled comparisons.	Standardize time/quality metrics; preregister outcome definitions; compare LLMs vs. baseline workflows; report prompt templates and interface features; test across educational levels and subjects.
RQ2: Automation and planning support	Use of LLMs to automate or assist sub-tasks (drafting plans, generating rubrics/assessments, feedback) and to streamline planning workflows; reported improvements vary by task and context.	Task-level automation indicators (e.g., completed steps, workflow time saved, self-reported extent of automation) and qualitative reports of usefulness/usability for planning.	Ambiguous definition of “automation” vs. “assistance”; limited reporting of negative/neutral outcomes; confounding from platform design and training; fast-changing tool capabilities.	Operationalize automation vs. assistance; evaluate usability and teacher adoption; include neutral/negative outcomes; separate model vs. platform effects; conduct longitudinal and classroom-based evaluations.

Insufficiently detailed instructions can lead LLMs to omit essential pedagogical elements. Fourteen studies ($n = 14$) warned that outputs may exclude learning objectives, contextualized examples, or formative-assessment components (Abreu et al., 2024; Albadarin et al., 2024; Al-Mughairi and Bhaskar, 2024; Gasaymeh and AlMohtadi, 2024; Isiaku et al., 2024; Ivanović, 2023; Jyothy et al., 2024; Lee and Song, 2024; Magalhães Araujo and Cruz-Correia, 2024; Sajadi et al., 2024; Shahzad et al., 2025; Song et al., 2024; Veras et al., 2024; Winder et al., 2024). Without careful manual validation, generated materials may lack internal coherence, lose scope, or overlook critical competencies. This challenge underscores the value of verification protocols, pedagogical checklists, and reusable prompt templates that systematically cover intended curriculum components.

Large language models frequently lack user-friendly interfaces that facilitate their integration into everyday teaching practice. Four studies (Magalhães Araujo and Cruz-Correia, 2024; Shahzad et al., 2025; Veras et al., 2024; Winder et al., 2024) highlighted a sole reliance on text boxes and prompts, the absence of configurable templates, and the lack of immediate feedback. These technical barriers force educators to toggle between platforms and learn syntax and commands, slowing the pedagogical workflow and potentially discouraging sustained use. To overcome this obstacle, it is essential to design intuitive graphical dashboards, pre-configured templates by subject and level, and keyboard shortcuts that streamline recurring tasks. In addition, incorporating real-time validations, such as suggestions and previews, would enhance both efficiency and user experience.

The assessment with ROBINS-I and CASP revealed uncertainty in critical domains that compromise the robustness of the findings. In the two quasi-experimental studies (Gasaymeh and AlMohtadi, 2024; Winder et al., 2024), five of the seven domains were rated as having unclear risk, whereas concerns about recruitment and reflexivity emerged in the qualitative and mixed-methods studies (Abreu et al., 2024; Al-Mughairi and Bhaskar, 2024; Isiaku et al., 2024; Ivanović, 2023; Jyothy et al., 2024; Lee and Song, 2024; Ma et al., 2023; Magalhães Araujo and Cruz-Correia, 2024; Sajadi et al., 2024; Song et al., 2024; Veras et al., 2024). The three reviews (Albadarin et al., 2024; Shahzad et al., 2025; Wang et al., 2024) displayed uncertainty in the presentation of results. These methodological weaknesses hinder

generalizability and may inflate the conclusions about effectiveness. More rigorous designs are required: randomization in quantitative studies, representative sampling in qualitative work, and complete transparency in outcome reporting to strengthen confidence in the evidence. Notably, for the quasi-experimental studies, five of seven ROBINS-I domains were rated as ‘unclear’, so claims about magnitude should be interpreted cautiously.

Time reduction and improvement in perceived quality

Across the primary studies, LLM-supported workflows were associated with reported reductions in the time required to generate educational materials or produce feedback for specific tasks, alongside perceived improvements in the clarity or usefulness of generated resources (e.g., Abreu et al., 2024; Sajadi et al., 2024; Winder et al., 2024). However, the magnitude of time savings varied across tasks and contexts and was not measured consistently; objective time measurements were reported in only a subset of studies, whereas others relied on self-report, perceived productivity, or qualitative accounts (Table 5). Several studies also noted limitations such as lack of depth, occasional inaccuracies, and the need for iterative prompting and human review to ensure alignment with learning objectives and context.

Task automation and planning efficiency

Across studies, LLMs were used to automate or assist sub-tasks such as question generation, rubric drafting, and first-pass feedback, with planning-efficiency indicators operationalized as steps eliminated or time saved (e.g., Gasaymeh and AlMohtadi, 2024; Sajadi et al., 2024; Winder et al., 2024). Reported automation/assistance estimates varied substantially across tasks and contexts, and heterogeneity of measures (self-reported percentages versus task-based indicators such as time or steps saved) limits direct comparability and does not support a single pooled estimate (Table 5). Qualitative evidence further suggests that workflow support (templates, iterative review, and integration

with existing planning routines) conditions whether teachers experience net time savings or added overhead.

Practical limitations based on the findings

Although the reviewed studies suggest that LLMs can support instructional planning and reduce time spent on content preparation, several practical limitations emerge from the evidence. (1) The frequent omission of key pedagogical elements and the sensitivity of outputs to minor prompt changes indicate that LLM-generated materials require systematic human review, domain verification, and alignment with learning objectives and assessment criteria before classroom use. In practice, this favors a human-in-the-loop workflow (draft → verify → adapt) rather than fully autonomous generation. (2) The magnitude of reported benefits is strongly conditioned by implementation factors that are often under-specified in the primary studies: teachers' prompt literacy, the availability of reusable prompt templates and exemplars, and user-friendly interfaces that integrate with existing lesson-planning routines. Where these supports are limited, time savings may be reduced or offset by additional effort spent on iterative prompting, editing, and troubleshooting. (3) Adoption is constrained by institutional and contextual requirements, including model access and cost, data privacy and compliance (especially when student data are involved), and governance policies regarding academic integrity and acceptable use. These constraints can restrict which LLMs are deployable, what inputs can be provided, and whether outputs can be integrated into official instructional planning systems.

Taken together, these factors limit the generalizability of reported effects and underscore that implementation readiness is as critical as model capability.

Implications for underrepresented educational contexts

Because most included studies were conducted in higher-education settings and often in comparatively well-resourced environments, implications for pre-tertiary/K-12, rural, low-resource, and multilingual contexts remain under-evidenced. Nonetheless, three cautious implications follow. (1) Access conditions (connectivity, device availability, subscription costs, and institutional procurement) may determine whether LLM use reduces workload or creates additional friction, particularly where teachers already face constraints in technology integration (Dotong et al., 2016; Jhurree, 2005). (2) Privacy, safeguarding, and policy requirements in school systems may restrict the use of consumer chat interfaces and increase the importance of governed deployments and data-minimization practices. (3) Language and cultural variation can affect perceived quality and fairness of generated materials, implying a need for locally validated templates, teacher professional development, and evaluation designs that explicitly test equity-relevant outcomes. These considerations reinforce the need for longitudinal, classroom-based studies in underrepresented contexts and for richer reporting of contextual variables that shape feasibility and impact.

Limitations of this review

This review should be interpreted considering several limitations.

- 1 Evidence base remains relatively small and heterogeneous, with substantial variation in study design, implementation settings, and outcome definitions, which limits direct comparability across studies.
- 2 Most included studies focus on higher education contexts and short-term deployments, constraining generalizability to other levels and to long-term adoption.
- 3 Several reported outcomes — time savings and perceived quality — are frequently measured via self-report or perception-based instruments rather than standardized, objective metrics, which increases uncertainty about effect magnitude.
- 4 Our synthesis includes both primary empirical studies and secondary evidence syntheses; while this broadens coverage, it also requires cautious interpretation because reviews do not provide equivalent empirical effect estimates.
- 5 Database coverage was limited to Scopus, ACM Digital Library, and Dimensions; consequently, education-specialist indexing services (e.g., ERIC, PsycINFO, and Web of Science education collections) were not searched, which may underrepresent mainstream education journals and pre-tertiary evidence.
- 6 Because our research questions and keyword clusters emphasize efficiency, time, and automation, there is a potential selection bias toward studies framed around productivity gains; relevant work reporting neutral or negative impacts — or focusing on other pedagogical outcomes without those terms — may be underrepresented.

Temporal limitations

All included studies evaluated models and interfaces available between 2023 and early 2025. Because LLM capabilities and deployment patterns — context-window sizes, retrieval features, safety layers, integration into LMS/workflows, and institutional policies — are evolving rapidly, our synthesis should be interpreted as a snapshot of that period rather than a stable, timeless description of “LLMs in education.” Accordingly, limitations discussed in this review are framed as constraints of the models and platforms used in the reviewed studies.

Future research directions

Future work would benefit from (1) standardized and transparent reporting of outcome measures (including objective time-use metrics and validated quality rubrics), (2) stronger comparative designs (controlled field studies, quasi-experiments, and replication studies across institutions), (3) broader coverage of educational levels, regions, and resource-constrained settings, and (4) implementation research that disentangles model-level effects from platform/workflow factors — prompting support, interface design, and teacher training. In addition, studies should more consistently evaluate governance issues

such as privacy, academic integrity, bias mitigation, and cost, to support responsible and scalable deployment.

Conclusion

Regarding RQ1, the reviewed studies suggest that using LLMs may reduce the time teachers devote to generating educational materials and may improve perceived quality in terms of clarity, coherence, and appropriateness. However, reported time-related benefits varied substantially across contexts and measurement approaches, and many outcomes relied on self-report rather than objective time-on-task measures. As for RQ2, the evidence indicates that LLMs can support automation or task assistance for pedagogical planning activities — lesson planning, schedules, rubrics, and classroom routines, potentially enabling more time for higher-value teacher–student interaction. Nevertheless, effectiveness depends heavily on prompt quality, implementation design, and the degree of technological integration within institutional settings.

Thus, applying LLMs in educational environments may offer a balance between efficiency and perceived quality enhancement; however, implementation also reveals tensions between potential benefits and well-documented risks. On the one hand, several studies reported reductions in time devoted to drafting materials and the automation or assistance of routine tasks, potentially freeing teachers to focus on pedagogical design and individualized support. On the other hand, the non-reproducibility of LLM outputs and incomplete content coverage call into question the reliability of generated materials and highlight the need for verification and human oversight.

User experience emerges as a decisive factor for the effective integration of LLMs into teaching practice. Beyond training teachers in prompt writing, it is essential to offer intuitive platforms that embed visual interfaces within learning-management systems. Configurable templates streamlined workflows that require only a few steps, and real-time feedback are strategies that minimize the learning curve and foster sustained adoption. This user-centered approach reduces technical friction and allows instructors to take advantage of LLMs without losing sight of their pedagogical role.

Taken together, within a limited and heterogeneous evidence base, the findings of this systematic review converge on four core themes: (1) the potential of LLMs to reduce teachers' workload and support the perceived quality of educational materials; (2) the technical and training conditions necessary for effective adoption, including instruction in prompt engineering and the design of user-friendly interfaces; (3) the urgency of establishing standardized metrics and rigorous validation mechanisms that ensure the reliability of generated content; and (4) the tension between the high potential of these tools and the risks associated with non-reproducibility, incomplete coverage, and the absence of clear ethical protocols. Given the small number of studies and the prevalence of unclear risk of bias in several domains, these conclusions should be viewed as indicative trends rather than definitive effect estimates. Recognizing and balancing these factors is fundamental to the sustainable and responsible integration of LLMs in education. These core themes respond directly to the research questions posed, providing a comprehensive perspective on efficiency, perceived quality, and

the automation of teachers' work when LLMs are incorporated into educational settings.

The findings raise a critical issue for the integration of Large Language Models (LLMs) into teaching: Can we trust the outputs they generate if adequate mechanisms for transparency and validation are not yet in place? As tools such as ChatGPT become increasingly accessible and widely used by both teachers and students, they are likely to become the primary source for designing instructional materials and resolving pedagogical queries — much as web search engines are today. This early adoption in the absence of clear validation and verification (V&V) protocols, turns the use of LLMs into an academic challenge: it is essential to establish quality standards, prompt traceability, and output audits before generated materials are routinely incorporated into the curriculum.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding author/s.

Author contributions

GL: Validation, Conceptualization, Methodology, Writing – original draft. RM-A: Validation, Formal analysis, Project administration, Writing – review & editing. ED: Supervision, Writing – review & editing, Validation. FF: Visualization, Writing – review & editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Acknowledgments

GL gratefully acknowledges SECIHTI for graduate scholarship No. 687862.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy,

including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or

claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/feduc.2026.1733861/full#supplementary-material>

References

- Abd-alrazaq, A., AlSaad, R., Alhuwail, D., Ahmed, A., Healy, P. M., Latifi, S., et al. (2023). Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med. Educ.* 9:e48291. doi: 10.2196/48291
- Abreu, A. A., Murimwa, G. Z., Farah, E., Stewart, J. W., Zhang, L., Rodriguez, J., et al. (2024). Enhancing readability of online patient-facing content: the role of AI chatbots in improving cancer information accessibility. *J. Natl. Compr. Cancer Netw.* 22:e237334. doi: 10.6004/jnccn.2023.7334
- Acerbi, A., and Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proc. Natl. Acad. Sci. USA* 120:e2313790120. doi: 10.1073/pnas.2313790120
- Albadarin, Y., Saqr, M., Pope, N., and Tukiainen, M. (2024). A systematic literature review of empirical research on ChatGPT in education. *Discov. Educ.* 3:60. doi: 10.1007/s44217-024-00138-2
- Alier, M., Casañ, M. J., and Amo Filvà, D. (2024). "Smart learning applications: leveraging LLMs for contextualized and ethical educational technology" in Proceedings of TEEM 2023. Lecture notes in educational technology. eds. J. A. C. Gonçalves, J. L. S. M. Lima, J. P. Coelho, F. J. García-Peñalvo and A. García-Holgado (Singapore: Springer), 190–199.
- Al-Mughairi, H., and Bhaskar, P. (2024). Exploring the factors affecting the adoption of AI techniques in higher education: insights from teachers' perspectives on ChatGPT. *J. Res. Innov. Teach. Learn.* doi: 10.1108/JRIT-09-2023-0129
- Alqahtani, T., Badreldin, H. A., Alrashed, M., Alshaya, A. I., Alghamdi, S. S., Bin Saleh, K., et al. (2023). The emergent role of artificial intelligence, natural language processing, and large language models in higher education and research. *Res. Soc. Adm. Pharm.* 19, 1236–1242. doi: 10.1016/j.sapharm.2023.05.016
- Caines, A., Benedetto, L., Taslimipour, S., Davis, C., Gao, Y., Andersen, Ø., et al (2023) On the application of large language models for language teaching and assessment technology. [Epub ahead of print]
- Çapuk, S., and Kara, A. (2015). A discussion of ICT integration within developed and developing world context from critical perspectives. *Procedia. Soc. Behav. Sci.* 191, 56–62. doi: 10.1016/j.sbspro.2015.04.411
- Chen, J., Liu, Z., Huang, X., Wu, C., Liu, Q., Jiang, G., et al. (2024). When large language models meet personalization: perspectives of challenges and opportunities. *World Wide Web* 27:42. doi: 10.1007/s11280-024-01276-1
- Critical Appraisal Skills Programme (2022) CASP systematic review checklist. Available online at: https://casp-uk.net/casp-checklists/CASP-Systematic-Review-checklist_2022.pdf (Accessed January 11, 2025)
- Dotong, C. I., De Castro, E. L., Dolot, J. A., and Prenda, M. T. B. (2016). Barriers for educational technology integration in contemporary classroom environment. *Asia Pac. J. Educ. Arts Sci.* 3, 13–20.
- Eriksen, M. B., and Frandsen, T. F. (2018). The impact of patient, intervention, comparison, outcome (PICO) as a search strategy tool on literature search quality: a systematic review. *J. Med. Libr. Assoc.* 106, 420–431. doi: 10.5195/jmla.2018.345
- Fernández-Batanero, J. M., Montenegro-Rueda, M., Fernández-Cerero, J., and García-Martínez, I. (2022). Digital competences for teacher professional development: systematic review. *Eur. J. Teach. Educ.* 45, 513–531. doi: 10.1080/02619768.2020.1827389
- Gan, W, Qi, Z, Wu, J, and Lin, JCW (2023) Large language models in education: vision and opportunities. In: *IEEE international conference on big data (BigData 2023)*, 15–18 Dec, Sorrento, Italy
- Gasaymeh, A.-M. M., and AlMohtadi, R. M. (2024). The effect of flipped interactive learning (FIL) based on ChatGPT on students' skills in a large programming class. *Int. J. Inf. Educ. Technol.* 14, 1516–1522. doi: 10.18178/ijiet.2024.14.11.2182
- Hadi, MU, Al-Tashi, Q, Qureshi, R, Shah, A, Muneer, A, Irfan, M, et al (2024) A survey on large language models: applications, challenges, limitations, and practical usage. [Epub ahead of print] doi: 10.36227/techrxiv.23589741.v1
- Huang, X., Zou, D., Cheng, G., Chen, X., and Xie, H. (2023). Trends, research issues and applications of artificial intelligence in language education. *Educ. Technol. Soc.* 26, 112–131. doi: 10.30191/ETS.202301_26(1).0009
- Isiaku, L., Kwala, A. F., Sambo, K. U., and Isiaku, H. H. (2024). Academic evolution in the age of ChatGPT: an in-depth qualitative exploration of its influence on research, learning, and ethics in higher education. *J. Univ. Teach. Learn. Pract.* 21. doi: 10.53761/7egat807
- Ivanović, I. (2023). Can AI-assisted essay assessment support teachers? A cross-sectional mixed-methods research conducted at the University of Montenegro. *Ann. Istr. Mediter. Stud. Ser. Hist. Sociol.* 33, 571–590. doi: 10.19233/ASHS.2023.30
- Jeon, J., and Lee, S. (2023). Large language models in education: a focus on the complementary relationship between human teachers and ChatGPT. *Educ. Inf. Technol.* 28, 15873–15892. doi: 10.1007/s10639-023-11834-1
- Jhurree, V. (2005). Technology integration in education in developing countries: guidelines to policy makers. *Int. Educ. J.* 6, 467–483.
- Jyothy, S. N., Kolil, V. K., Raman, R., and Achuthan, K. (2024). Exploring large language models as an integrated tool for learning, teaching, and research through the Fogg behavior model: a comprehensive mixed-methods analysis. *Cogent Engin.* 11:2353494. doi: 10.1080/23311916.2024.2353494
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* 103:102274. doi: 10.1016/j.lindif.2023.102274
- Lee, J., Hicke, Y., Yu, R., Brooks, C., and Kizilcec, R. F. (2024). The life cycle of large language models in education: a framework for understanding sources of bias. *Br. J. Educ. Technol.* 55, 1982–2002. doi: 10.1111/bjet.13505
- Lee, S., and Song, K.-S. (2024). Teachers' and students' perceptions of AI-generated concept explanations: implications for integrating generative AI in computer science education. *Comput. Educ.* 7:100283. doi: 10.1016/j.caeai.2024.100283
- Li, J., Tang, T., Zhao, W. X., Nie, J.-Y., and Wen, J.-R. (2024). Pre-trained language models for text generation: a survey. *ACM Comput. Surv.* 56, 1–39. doi: 10.1145/3649449
- Liu, F, Wang, Y, Feng, Q, Zhu, L, and Li, G (2024) Optimizing e-learning environments: leveraging large language models for personalized education pathways. In: Y Kuang, et al. (eds) Proceedings of the 2024 5th international conference on education, knowledge and information management (ICEKIM 2024). Atlantis Highlights in Computer Sciences, vol. 22. Amsterdam: Atlantis Press, pp. 812–817
- Ma, Q, Shen, H, Koedinger, K, and Wu, ST 2023 How to teach programming in the AI era? Using LLMs as a teachable agent for debugging [Epub ahead of print] doi: 10.48550/arXiv.2310.05292
- MacKinnon, P. C., and MacKinnon, G. (2013). Technology integration in developing countries: a case study of higher education in Jamaica. *Int. J. Technol. Knowledge. Soc.* 9, 51–59. doi: 10.18848/1832-3669/CGP/v09i01/56344
- Magalhães Araujo, S., and Cruz-Correia, R. (2024). Incorporating ChatGPT in medical informatics education: mixed methods study on student perceptions and experiential integration proposals. *JMIR Med. Educ.* 10:e51151. doi: 10.2196/51151
- Makgato, M. (2014). Challenges contributing to poor integration of educational technology at some schools in South Africa. *Mediterr. J. Soc. Sci.* 5, 1285–1292. doi: 10.5901/mjss.2014.v5n20p1285
- Mansur, H., Utama, A. H., Mohd Yasin, M. H., Sari, N. P., Jamaludin, K. A., and Pinandhita, F. (2023). Development of inclusive education learning design in the era of society 5.0. *Soc. Sci.* 12:35. doi: 10.3390/socsci12010035
- Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., et al. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Comput. Educ.* 6:100199. doi: 10.1016/j.caeai.2023.100199

- Meyer, J. G., Urbanowicz, R. J., Martin, P. C. N., Li, R., Peng, P. C., Bright, T. J., et al. (2023). ChatGPT and large language models in academia: opportunities and challenges. *BioData Min.* 16:20. doi: 10.1186/s13040-023-00339-9
- Milano, S., McGrane, J. A., and Leonelli, S. (2023). Large language models challenge the future of higher education. *Nat. Mach. Intell.* 5, 333–334. doi: 10.1038/s42256-023-00644-2
- Ningsih, W., and Lahby, M. (2025). “The potential of GPT in education: opportunities, limitations, and recommendations for adaptive learning” in Empowering digital education with ChatGPT: From theoretical to practical applications. ed. M. Lahby. 1st ed (Boca Raton, FL: CRC Press), 247–263.
- Onyema, E. M. (2019). Integration of emerging technologies in teaching and learning process in Nigeria: the challenges. *Central Asian J. Math. Theory Comput. Sci.* 1, 36–39.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372:n71. doi: 10.1136/bmj.n71
- Park, M., Kim, S., Lee, S., Kwon, S., and Kim, K. (2024). “Empowering personalized learning through a conversation-based tutoring system with student modeling” in Extended abstracts of the CHI conference on human factors in computing systems (CHI EA '24) (Honolulu, HI, USA. New York, NY: ACM), 1–10.
- Peláez-Sánchez, I. C., Velarde-Camaqui, D., and Glasserman-Morales, L. D. (2024). The impact of large language models on higher education: exploring the connection between AI and education 4.0. *Front. Educ.* 9:1392091. doi: 10.3389/feduc.2024.1392091
- Ramorola, M. Z. (2013). Challenge of effective technology integration into teaching and learning. *Afr. Educ. Rev.* 10, 654–670. doi: 10.1080/18146627.2013.853559
- Roshanaei, M. (2024). Towards best practices for mitigating artificial intelligence implicit bias in shaping diversity, inclusion and equity in higher education. *Educ. Inf. Technol.* 29, 18959–18984. doi: 10.1007/s10639-024-12605-2
- Safranek, C. W., Sidamon-Eristoff, A. E., Gilson, A., and Chartash, D. (2023). The role of large language models in medical education: applications and implications. *JMIR Med. Educ.* 9:e50945. doi: 10.2196/50945
- Sajadi, S., Huerta, M., Ryan, O., and Drinkwater, K. (2024). Harnessing generative AI to enhance feedback quality in peer evaluations within project-based learning contexts. *Int. J. Eng. Educ.* 40, 998–1012.
- Shahzad, T., Mazhar, T., Tariq, M. U., Ahmad, W., Ouahada, K., and Hamam, H. (2025). A comprehensive review of large language models: issues and solutions in learning environments. *Discov. Sustain.* 6:27. doi: 10.1007/s43621-025-00815-8
- Sharma, S., Mittal, P., Kumar, M., and Bhardwaj, V. (2025). The role of large language models in personalized learning: a systematic review of educational impact. *Discov. Sustain.* 6:243. doi: 10.1007/s43621-025-01094-z
- Song, X., Zhang, J., Yan, P., Hahn, J., Kruger, U., Mohamed, H., et al. (2024). Integrating AI in college education: positive yet mixed experiences with ChatGPT. *Meta-Radiology* 2:100113. doi: 10.48550/arXiv.2407.05810
- Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., et al. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 355:i4919. doi: 10.1136/bmj.i4919
- Veras, M., Dyer, J.-O., Shannon, H., Bogie, B. J. M., Ronney, M., Sekhon, H., et al. (2024). A mixed methods crossover randomized controlled trial exploring the experiences, perceptions, and usability of artificial intelligence (ChatGPT) in health sciences education. *Digit. Health* 10:20552076241298485. doi: 10.1177/2055207624129848
- Walter, Y. (2024). Managing the race to the moon: global policy and governance in artificial intelligence regulation—a contemporary overview and an analysis of socioeconomic consequences. *Discov. Artif. Intell.* 4:14. doi: 10.1007/s44163-024-00109-4
- Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., et al. 2024 Large language models for education: a survey and outlook [Epub ahead of print] doi: 10.48550/arXiv.2403.18105
- Winder, G., Bass, S., Schiele, D., and Buchner, J. (2024). Using large language models for content creation impacts online learning evaluation outcomes. *Int. J. E-Learn.* 24, 305–318. doi: 10.70725/423664moqcrd
- Xu, H., Gan, W., Qi, Z., Wu, J., and Yu, P. S. 2024 Large language models for education: a survey [Epub ahead of print]. doi: 10.48550/arXiv.2405.13001
- Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., et al. (2024). Practical and ethical challenges of large language models in education: a systematic scoping review. *Br. J. Educ. Technol.* 55, 90–112. doi: 10.1111/bjet.13370

Appendix A

Database search strategies

To improve transparency and reproducibility, we report the database-specific search strategies used in this review. Scopus was selected for its broad multidisciplinary coverage, ACM Digital Library to capture computer-science and educational-technology venues, and Dimensions to broaden coverage across publishers. We acknowledge that education-focused indexes such as ERIC, PsycINFO and Web of Science education collections were not searched, which may underrepresent mainstream education research; this is treated as a limitation.